

Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie Wydział Informatyki, Elektroniki i Telekomunikacji Instytut Informatyki

Języki formalne i kompilatory

Projekt zaliczeniowy

Translator języka LATEX do HTML

Bartosz Kordek Grzegorz Zacharski

Informatyka Studia niestacjonarne

Kraków, 2021

Spis treści

1	$\mathbf{W}\mathbf{s}$ 1	tęp	2	
	1.1	Pliki tworzące projekt	2	
	1.2	Uruchamianie	2	
2	Obs	sługiwany zakres translatora	4	
	2.1	Tekst	4	
	2.2	Formatowanie tekstu	4	
	2.3	Tabela	4	
	2.4	Wyliczenie	5	
	2.5	Grafika	7	
	2.6	Hiperłącze	1	
	2.7	Sekcja, podsekcja, podpodsekcja	5	
3	Gramatyka			
	3.1	Lexer	6	
	3.2	Parser	8	
		3.2.1 Obsługa tekstu	S	
		3.2.2 Formatowanie tekstu	6	
		3.2.3 Tabela	11	
			12	
		3.2.5 Grafika	13	
		3.2.6 Hiperłącze	14	
4	Uza	asadnienie wyboru generatora	15	
5	Nap	potkane problemy	16	
	5.1	Przejście do nowej linii	16	
	E 9	Translacia tabali	16	

Wstęp

Celem projektu było zaimplementowanie translatora języka LATEX do HTML. W tym celu skorzystaliśmy z gotowej biblioteki PLY [3] do skanowania i parsowania napisanej w języku Python. Kod odpowiadający za wykonanie translacji podzieliliśmy na dwie części: lekser (skanowanie) i parser (parsowanie).

Do wykonania projektu skorzystaliśmy również z oficjalnej dokumentacji języka I^ATEX [2] oraz języka HTML [1].

1.1 Pliki tworzące projekt

Translator został zaimplementowany w języku Python 3. Projekt składa się z następujących plików:

- main.py plik wejściowy programu, odpowiada za uruchomienie interpretera, wyświetlanie pomocniczych informacji w trybie debugowania oraz za uruchomienie odpowiedniego trybu działania programu
- theLexer.py lexer, który ma za zadanie zwrócić tokeny po przeanalizowaniu kodu dokumentu latex
- the Parser.py - parser, który ma za zadanie stworzyć abstrakcyjne drzewo syntaktyczne oraz zwrócie kod html
- example.tex przykładowy plik testowy, który zostanie przetłumaczony do języka html

1.2 Uruchamianie

W celu uruchomienia tranlatora trzeba mieć zainstalowany interpreter języka Python 3 oraz w folderze projektu wpisać w terminalu komendę pokazaną na listingu 1.1:

python3 main.py -i example -o example

Listing 1.1: Uruchomienie

Parametr i przyjmuje nazwę pliku do translacji, natomiast paramentr o przyjmuje nazwę pliki html, który zostanie stworzony podczas działania programu.

Obsługiwany zakres translatora

2.1 Tekst

Nie wyszczególniono specjalnych wymagań, które musi spełniać zwykły tekst w języku IATEX , aby został przetłumaczony na język HTML.

2.2 Formatowanie tekstu

Translator obsługuje następujące znaczniki formatu LaTeX odpowiadające na formatowanie tekstu:

- \underline podkreślenie
- \centerline wyśrodkowanie
- \title tytuł

2.3 Tabela

W celu przetłumaczenia tabeli z języka L^AT_EX na język HTML tabela powinna spełniać powniższe założenia:

- musi być zadeklarowana poleceniem $\begin{tabular} tabular \end{tabular}$ oraz zakończona $\end{tabular}$
- po deklaracji $\begin{tabular}$ musi zostać od razu zadeklarowane obramowanie w
g standardu LATEX , np. $\begin{tabular}{c} c c c \end{tabular}$

- \bullet tabele z obramowaniem muszą zostać zadeklarowane w
g $\begin{tabular}\{\mid c\mid c\mid c\mid\}$ ilość kolumn dowolna
- \bullet tabele bez obramowania muszą zostać zadeklarowane w
g $\begin{tabular}\{c\ c\ c\}$ ilość kolumn dowolna

2.4 Wyliczenie

Program obsługuje wyliczenie uporządkowane i nieuporządkowane:

- \begin{enumerate} uporządkowane

2.5 Grafika

Grafikę w tekście możemy zadeklarować wyłącznie używając polecenia $\include graphics$ podając ściezkę do zdjęcia np: $\include graphics \{ corgi.jpq \}$.

2.6 Hiperłącze

Hiperłącze w deklarujemy wyłącznie poprzez użycie \url np. $\url\{https://github.com/bartoszkordek/AGHJezyki-formalne-i-kompilatory\}$

2.7 Sekcja, podsekcja, podpodsekcja

Sekcje, podsekcje oraz podpodsekcje deklarujemy w ten sam sposób jak w języku LATEX np. $\subsection{Przykladowa podsekcja}.$

Gramatyka

3.1 Lexer

Na listingu 3.1 zostały przedstawione tokeny obsługiwane przez nasz translator.

```
tokens = (
           'AUTHOR',
           'BEGIN_DOCUMENT',
           'BEGIN_OLIST',
           'BEGIN_ULIST',
           'BEGIN_TABULAR',
           'BOLD',
           'CENTERLINE',
           'CHAPTER',
           'COLUMN_DIVIDER',
10
           'COLUMN_PATTERN_BORDERLESS',
           'COLUMN_PATTERN_BORDERED',
           'DOCUMENTCLASS',
           'END_DOCUMENT',
14
           'END_OLIST',
15
           'END_ULIST',
16
           'END_TABULAR',
17
           'GRAPHICS_PATH',
           'HLINE',
19
           'INCLUDE_GRAPHICS',
20
           'ITALIC',
           'ITEM',
           'LBRACE',
23
           'NEW_LINE',
           'NULL',
           'PARAGRAPH',
26
           'RBRACE',
           'ROW_END',
           'SECTION',
```

Listing 3.1: Tokeny

Tokeny na listingu 3.1 odpowiadają następującym wyrażeniom regularnym przedstawionym na listingu 3.2:

```
t_AUTHOR = r' \setminus author'
       t_BEGIN_DOCUMENT = r'\\begin\{document\}'
       t_BEGIN_OLIST = r'\\begin\{enumerate\}'
       t_BEGIN_ULIST = r'\\begin\{itemize\}'
       t_BEGIN_TABULAR = r'\\begin\{tabular\}'
       t_BOLD = r' \setminus textbf'
       t_CENTERLINE = r'\\centerline'
       t_CHAPTER = r'\\chapter'
       t_COLUMN_DIVIDER = r'&'
       t_COLUMN_PATTERN_BORDERLESS = r'\{[lcr](\s[lcr])*\}'
       t_COLUMN_PATTERN_BORDERED = r'\setminus\{(\backslash |\backslash s[lcr]\backslash s) + \backslash |\backslash \}'
       t_DOCUMENTCLASS = r'\\documentclass.*'
       t_END_DOCUMENT = r'\\end\{document\}'
13
       t_END_OLIST = r'\\end\{enumerate\}'
14
       t_END_ULIST = r'\\end\{itemize\}'
       t_END_TABULAR = r'\\end\{tabular\}'
       t_GRAPHICS_PATH = r'\\graphicspath'
17
       t_INCLUDE_GRAPHICS = r'\\includegraphics'
18
       t_ITALIC = r'\\textit'
19
       t_{ITEM} = r' \setminus item'
20
       t_LBRACE = r' \setminus \{'\}
       t_NEW_LINE = r'\\newline'
22
       t_NULL = r' \setminus 0'
       t_PARAGRAPH = r'\\paragraph'
       t_RBRACE = r' \
       t_ROW_END = r' / / / '
26
       t_SECTION = r'\\section'
       t_SUBSECTION = r'\\subsection'
28
       t_SUBSUBSECTION = r'\\subsubsection'
29
       t_{TEXT} = r'[\w\d\.,!?@#/\'\"<>\(\)\-+=\/^\*:;|\[\]]+'
       t_TITLE = r'\\title'
31
       t_UNDERLINE = r'\\underline'
32
       t_URL = r'\\url'
```

```
t_USE_PACKAGE = r'\\usepackage.*'
```

Listing 3.2: Wyrażenia regularne

Definicje tokenów nowej linii, komentarza, błędów oraz tokenów ignorowanych zostały przedstawione na listingu 3.3.

```
def t_newline(self, t):
    r'\n+'
    t.lexer.lineno += t.value.count("\n")

def t_comment(self, t):
    r'%.*\n'
    pass

def t_hline(self, t):
    r'\\hline'
    pass

t_ignore = ' '

def t_error(self, t):
    print("Illegal character '%s'" % t.value[0])
    t.lexer.skip(1)
```

Listing 3.3: Pozostałe tokeny

Należy zauważyć, że cała zawartość między znacznikiem "%"w formacie L^AT_EX, odpowiadającemu początkowi komentarza, do nowej linii, zostaje pominieta.

3.2 Parser

Każdy z dokumentów HTML'a musi zawierać znaczniki typowe dla tego formatu, których nie znajdziemy w dokumentach L^AT_EX. Są to: *html*, *head* oraz *body*. Jako odpowienik sekcji *head* przyjęliśmy sekcję preambuly w której znajdują się między innymi takie polecenia jak *usepackage*. Natomiast jak odpowiednik sekcji *body* odpowiada fragment kodu L^AT_EX objęty poleceniami \ *begin*{ document} oraz \ end{ document}.

```
def p_body(self, p):
    'body : BEGIN_DOCUMENT expression END_DOCUMENT'
    p[0] = '\n<body>\n' + p[2] + '\n</body>'
```

Listing 3.4: Preambuła

3.2.1 Obsługa tekstu

Listing 3.5: Tekst

3.2.2 Formatowanie tekstu

Nasz translator umożliwia pogrubienie, kursywę, podkreślenie, wyśrodkowanie tekstu oraz utworzenie paragrafu. Możliwe jest mieszanie stylów formatowania, np. pogrubienie z podkreśleniem. Według nowych zaleceń, do pogrubienia tekstu w HTML'u powinno się stosować znacznik $\langle stronq \rangle$, a do kursywy $\langle em \rangle$.

```
def p_expression_bold(self, p):
          ''', expression : BOLD LBRACE expression RBRACE expression
                         | BOLD LBRACE expression RBRACE'''
          if len(p) == 6:
              p[0] = '<strong>' + p[3] + '</strong>' + p[5]
          else:
              p[0] = '<strong>' + p[3] + '</strong>'
      def p_expression_italic(self, p):
9
          ''', expression : ITALIC LBRACE expression RBRACE expression
                         | ITALIC LBRACE expression RBRACE'''
          if len(p) == 6:
              p[0] = '<em>' + p[3] + '</em>' + p[5]
13
          else:
14
              p[0] = '<em>' + p[3] + '</em>'
      def p_expression_underline(self, p):
17
          '''expression : UNDERLINE LBRACE expression RBRACE expression
                         | UNDERLINE LBRACE expression RBRACE'''
19
          if len(p) == 6:
20
              p[0] = (u) + p[3] + (u) + p[5]
```

```
else:
22
              p[0] = '<u>' + p[3] + '</u>'
24
      def p_expression_centerline(self, p):
25
          ''', expression : CENTERLINE LBRACE expression RBRACE expression
                         | CENTERLINE LBRACE expression RBRACE'''
          if len(p) == 6:
28
              p[0] = '<center>' + p[3] + '</center>' + p[5]
          else:
30
              p[0] = '<center>' + p[3] + '</center>'
31
32
      def p_expression_paragraph(self, p):
33
          '''expression : PARAGRAPH LBRACE expression RBRACE expression
34
                         | PARAGRAPH LBRACE expression RBRACE'''
35
          if len(p) == 6:
36
              p[0] = '' + p[3] + '' + p[5]
              p[0] = '' + p[3] + ''
39
```

Listing 3.6: Formatowanie - pogrubienie, kursywa, podkreślenie, wyśrodkowanie tekstu

Kolejną funkcjonalnością jest możliwość obsługi rozdziałów, sekcji i podsekcji parsując znaczniki *chapter*, *section*, *subsection* oraz *subsubsection*.

```
def p_expression_chapter(self, p):
          '''expression : CHAPTER LBRACE expression RBRACE expression
                         | CHAPTER LBRACE expression RBRACE'''
          if len(p) == 6:
              p[0] = (h1) + p[3] + (h1) + p[5]
          else:
              p[0] = '<h1>' + p[3] + '</h1>'
      def p_expression_section(self, p):
          ''', expression : SECTION LBRACE expression RBRACE expression
                         | SECTION LBRACE expression RBRACE'''
          if len(p) == 6:
12
              p[0] = (h2) + p[3] + (h2) + p[5]
          else:
14
              p[0] = '<h2>' + p[3] + '</h2>'
      def p_expression_subsection(self, p):
17
          '''expression : SUBSECTION LBRACE expression RBRACE expression
18
                         | SUBSECTION LBRACE expression RBRACE'''
19
          if len(p) == 6:
              p[0] = \frac{(h3)}{p[3]} + \frac{(/h3)}{p[5]}
21
          else:
              p[0] = '<h3>' + p[3] + '</h3>'
23
```

Listing 3.7: Rozdziały, sekcje i podsekcje

Przejście do nowej linii (hard break) jest obsłużone poleceniem newline.

Listing 3.8: Nowa linia

Translator parsuje również znacznik title odpowiadający za utworzenie tytułu.

Listing 3.9: Tytuł

3.2.3 Tabela

Z obramowaniem

```
def p_expression_table_bordered(self, p):
         ''', expression : BEGIN_TABULAR COLUMN_PATTERN_BORDERED
    tablerowbordered END_TABULAR expression | BEGIN_TABULAR
    COLUMN_PATTERN_BORDERED tablerowbordered END_TABULAR'',
         if len(p) == 6:
             p[0] = '<table style="border: 1px solid black; border-collapse:
    collapse;">' + p[3] + '' + p[5]
         else:
             p[0] = '
    collapse;">' + p[3] + ''
     def p_tablerowbordered(self, p):
         '''tablerowbordered : tablecolumnbordered ROW_END tablerowbordered |
     tablecolumnbordered',',
         if len(p) == 4:
10
             p[0] = '\langle tr \rangle' + p[1] + '\langle tr \rangle' + p[3]
```

```
else:
    p[0] = '' + p[1] + ''

def p_tablecolumnbordered(self, p):
    '''tablecolumnbordered : expression COLUMN_DIVIDER

tablecolumnbordered | expression'''

if len(p) == 4:
    p[0] = '' + p[1] + '' + p[3]

else:
    p[0] = '' + p[1] + ''
```

Listing 3.10: Tabela z obramowaniem

Bez obramowania

```
def p_expression_table_borderless(self, p):
          '''expression : BEGIN_TABULAR COLUMN_PATTERN_BORDERLESS
     tablerowborderless END_TABULAR expression | BEGIN_TABULAR
     COLUMN_PATTERN_BORDERLESS tablerowborderless END_TABULAR'',
          if len(p) == 6:
3
              p[0] = '' + p[3] + '' + p[5]
          else:
              p[0] = '\langle table \rangle' + p[3] + '\langle table \rangle'
6
      def p_tablerowborderless(self, p):
          '', tablerowborderless : tablecolumnborderless ROW_END
     tablerowborderless | tablecolumnborderless',',
          if len(p) == 4:
              p[0] = '' + p[1] + '' + p[3]
          else:
              p[0] = '' + p[1] + ''
      def p_tablecolumnborderless(self, p):
          '', 'tablecolumnborderless : expression COLUMN_DIVIDER
16
     tablecolumnborderless | expression'''
          if len(p) == 4:
              p[0] = '' + p[1] + '' + p[3]
18
          else:
19
              p[0] = '' + p[1] + ''
```

Listing 3.11: Tabela bez obramowania

3.2.4 Wyliczenie

Konstrukcja parsera wyliczeń umożliwia wykonywanie zagnieżdzeń.

Uporządkowane

Listing 3.12: Wyliczenie uporządkowane

Nieuporządkowane

Listing 3.13: Wyliczenie nieuporządkowane

3.2.5 Grafika

Umieszcznie grafiki jest możliwe dzięki znacznikowi includegraphics w \LaTeX który jest parsowany na znacznik < img> w \LaTeX HTML, gdzie atrybut src stanowi ścieżka do pliku umieszczona w nawiasach wąsatych w dokumencie \LaTeX .

```
def p_expression_includegraphics(self, p):
          '''expression : INCLUDE_GRAPHICS TEXT LBRACE expression RBRACE
    expression
                        | INCLUDE_GRAPHICS LBRACE expression RBRACE expression
3
                        | INCLUDE_GRAPHICS LBRACE expression RBRACE'''
         if len(p) == 7:
              attributes = p[2][1:-1]
6
             p[0] = '''<img src="''' + p[4] + '''"''' + attributes + '''>'''
    + p[6]
         elif len(p) == 6:
              p[0] = '''<img src="''' + p[3] + ''''>''' + p[5]
9
         else:
             p[0] = '''<img src="''' + p[3] + ''''>'''
```

Listing 3.14: Grafika

3.2.6 Hiperłącze

Zamieszczenie hiperłącza w formacie LaTeXjest możliwe dzięki znacznikowi url, zawierającego w nawiasach wąsatych adres do strony. Parsowanny jest on na HTML'owy znacznik < a > z atrybutem href zawierającego adres.

Listing 3.15: Hiperłącze

Uzasadnienie wyboru generatora

Podstawowym wymaganiem w projekcie była konieczność zastowania generatora parserów i skanerów. Z zaproponowanych na zajęciach bibliotek (PLY, SLY oraz ANTLR) zdecydowaliśmy się na wybór biblioteki PLY [3], ze względu na jej popularność stosowania, dostępność i wysoką jakość dokumentacji. Ponadto chcieliśmy również zdobyć kolejne doświadczenie w implementacji kodu z wykorzystaniem języka Python. Na niekorzyść wykorzystania biblioteki SLY przemawiała niska jakość dokumentacji oraz brak przykładów.

Napotkane problemy

Podczas realizacji projektu napotkaliśmy następujące problemy do rozwiązania:

- przejście do nowej linii
- translacja tabeli

5.1 Przejście do nowej linii

W języku Late przejście do nowej linii jest możliwe dzięki następujący sposób: wykorzystując znacznik $\ensuremath{\backslash} newline$, podwójnemu backslashowi, oraz wykonaniem podwójnego prześcia do nowej linii za pomocą klawisza Enter. W rezultacie chcieliśmy otrzymać znacznik $\ensuremath{\langle} br\ensuremath{\rangle} w$ HTMLu. Nie napotkaliśmy większych problemów z translacją znacznika $\ensuremath{\rangle} newline$. Podwójny backslash również działał poprawnie dopóki nie zaimplementowaliśmy translacji tabeli, w której ten sam znacznik oznacza jej koniec. Ze względu na ten problem musieliśmy z niego zrezygnować. Problem nastąpił również z przejściem do nowej linii przy pomocy dwukrotnego wciśniecia klawisza Enter. W tym przypadku każde przejście do nowej linii było traktowane jako przejście pojedyncze, a podwójne było pomijane. W rezultacie oznaczało to, że nie mogliśmy otrzymać znacznika $\ensuremath{\langle} br\ensuremath{\rangle}$ w HTMLu. Po usunięciu definicji tokena $\ensuremath{\rangle}$ newline, znacznik $\ensuremath{\langle} br\ensuremath{\rangle}$ był produkowany, lecz program generował ostrzeżenie (brak znajomości tokena) w przypadku wystąpienia pojedynczego przejścia. Plik w formacie HTML był jednak poprawnie generowany. Ze względu na ostrzeżenie, pomysł został porzucony.

5.2 Translacja tabeli

Podstawowym problemem był opisany powyżej znacznik przejścia do nowej linii, który w przypadku tabeli oznacza jej koniec. Ponadto tabela jest na tyle złożoną sktrukturą, że nie jesteśmy w stanie zapewnić obsługi każdego z jej przypadków. Uprościliśmy więc translację do tabeli w pełni obramowanej lub nieobramowanej.

Listings

1.1	Uruchomienie	3
3.1	Tokeny	6
3.2	Wyrażenia regularne	7
3.3	Pozostałe tokeny	8
3.4	Preambuła	8
3.5	Tekst	9
3.6	Formatowanie - pogrubienie, kursywa, podkreślenie, wyśrodkowanie tekstu . .	9
3.7	Rozdziały, sekcje i podsekcje	10
3.8	Nowa linia	11
3.9	Tytuł	11
3.10	Tabela z obramowaniem	11
3.11	Tabela bez obramowania	12
3.12	Wyliczenie uporządkowane	13
3.13	Wyliczenie nieuporządkowane	13
3.14	Grafika	13
3.15	Hiperlacze	14

Bibliografia

- [1] HTML 5.2. W3C Recommendation. 2021. URL: https://www.w3.org/TR/html52/.
- [2] Overleaf. Official Documentation. 2021. URL: https://www.overleaf.com/learn.
- [3] PLY. Official Documentation. 2021. URL: https://ply.readthedocs.io/en/latest/.